Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W485–W491***
DOI: 10.1093/nar/gkh421

New challenges in gene expression data analysis and the extended GEPAS

Javier Herrero, Juan M. Vaquerizas, Fátima Al-Shahrour, Lucía Conde, Álvaro Mateos, Javier Santoyo Ramón Díaz-Uriarte and Joaquín Dopazo*

Bioinformatics Unit, Biotechnology Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro, 3, E-28029 Madrid, Spain

Received February 13, 2004; Revised and Accepted April 7, 2004

ABSTRACT

Since the first papers published in the late nineties, including, for the first time, a comprehensive analysis of microarray data, the number of questions that have been addressed through this technique have both increased and diversified. Initially, interest focussed on genes coexpressing across sets of experimental conditions, implying, essentially, the use of clustering techniques. Recently, however, interest has focussed more on finding genes differentially expressed among distinct classes of experiments, or correlated to diverse clinical outcomes, as well as in building predictors. In addition to this, the availability of accurate genomic data and the recent implementation of CGH arrays has made mapping expression and genomic data on the chromosomes possible. There is also a clear demand for methods that allow the automatic transfer of biological information to the results of microarray experiments. Different initiatives, such as the Gene Ontology (GO) consortium, pathways databases, protein functional motifs, etc., provide curated annotations for genes. Whereas many resources on the web focus mainly on clustering methods, GEPAS has evolved to cope with the aforementioned new challenges that have recently arisen in the field of microarray data analysis. The web-based pipeline for microarray gene expression data, GEPAS, is available at http://gepas.bioinfo.cnio.es.

INTRODUCTION

Gene expression analysis using microarray technology has opened up a wide range of possibilities for exploring the biology of cells and organisms. In the early days, interest was mainly focussed on the behaviour of genes across the experimental conditions studied (1); recently though, biomedical applications [e.g. (2–4)] have fuelled both the use of available technologies and the development and implementation of analytical tools. In terms of data analysis methodologies, it is implied that, in addition to clustering, there is high demand for efficient methods for class prediction, which would include the derivation of prognosis predictors, response to drugs or therapies, or any phenotype or genotype defined independently of the gene expression profile. The availability of accurate genome assemblies in public repositories such as the Ensembl (http://www.ensembl.org) or the NCBI (http://www.ncbi.nih.gov/Genomes/index.html) allows mapping expression data over the genome. Moreover, new microarray-related technologies, such as microarray-based comparative genomic hybridization [array CGH; (4)] introduce another dimension in the analysis: the possibility of obtaining precise mapping of copy number alterations in the genome. Perhaps one of the most demanded kinds of tools are those that can transfer biologically relevant information to microarray experiments. This information can be extracted either from free text (e.g. Medline abstracts) or from more or less curated repositories. The use of text mining techniques in studying the coherence of gene groups obtained from different methodologies has only recently been addressed (5–7), although its practical application still poses many drawbacks (8). Furthermore, availability to end users is often scarce. Gene Ontology [GO; (9)], which organizes information for molecular function, biological processes and cellular components for a number of different organisms, and KEGG (10), which includes a comprehensive description of different pathways, are among the most used curated repositories of information. Different tools, which generate tables correlating groups of genes to GO terms regarding biochemical and molecular functions, have been recently implemented [see (11–13) and the web page of the GO consortium http://www.geneontology.org]. Among them is FatiGO (13,14), which provides an appropriate statistical framework which takes into account the multiple-testing nature of the statistical contrast [an important fact often

*To whom correspondence should be addressed. Tel: +34 912 246 919; Fax: +34 912 246 972; Email: jdopazo@cnio.es

neglected (15)], and has the advantage of being integrated within the GEPAS (16) platform.

GEPAS constitutes an integrated web-based pipeline for the analysis of gene expression patterns where different methods can be used within an integrated interface, providing a user-friendly environment to end users. The way in which the methods are connected guide the user by suggesting all the available possibilities to continue with analysis. The use of methods is largely conditioned by its availability as programs. For example, the overabundance of tools focussing on cluster analysis has lead to a misuse of this methodology. In fact, some authors have specifically highlighted the inappropriate use of clustering for class comparison or class prediction (17). Given this, it is important to make appropriate methods available for dealing with different problems such as class prediction, gene selection, data mining of the results, etc. Since the original version (16) of our tool, GEPAS, the scope and the number of methods included have been increased significantly, in order to cope with new challenges that have arisen in the field of microarray data analysis.

SCOPE OF GEPAS

The original design of GEPAS (16), along with the new additions intends to cover as many experimental situations as possible and to respond to different scientific and clinical questions. With this goal in mind, we have implemented a series of tools (some of them publicly available and others developed by us) within a web-based pipeline of microarray data analysis. The following sections describe the architecture of the pipeline, the methods included in GEPAS and the new methods added since the first version in more depth.

A WEB-BASED PIPELINE FOR MICROARRAY DATA ANALYSIS

Figure 1 shows how the different modules of GEPAS are interconnected and can exchange data. Once the expression pattern data is introduced within the system, all the available modules can be used for analysis. The pipeline of microarray data analysis starts with the results of the quantification of the spots, corresponding to the hybridizations, provided by the program that processes the image obtained by the scanner. These intensities are assumed to be proportional to the amount of mRNAs corresponding to the probes in the microarray. Depending on the technology used [cDNA microarrays (18) or Affymetrix oligonucleotide arrays (19)], the mRNA amounts are measured as absolute values or as ratios with respect to another reference mRNA. Irrespective of this, the first necessary step is normalization. In this step, differences occurring for reasons other than those sought in the experiment (biases, local effects, differences in efficiency of the hybridization, etc.) are removed. The matrix of normalized gene expression values is then sent to the preprocessor (20), a module that carries out a series of operations which may be required (such as missing value imputation, filtering of 'flat' patterns, etc.) and at the same time acts as a central hub for distributing the data among the different methods implemented. Depending on the problem, data can be sent to modules for clustering, gene selection, class (or any phenotypic trait) prediction, genome mapping and data mining, thus responding to a significant number of data analysis requirements.

The efficiency of a modular package such as GEPAS lies in its degree of integration of the different data analysis tools. Users can cover a complete pipeline of data analysis in a transparent way, without the necessity of performing any reformatting operation. In addition, web-based tools guarantee real cross-platform capabilities. Client-server architecture

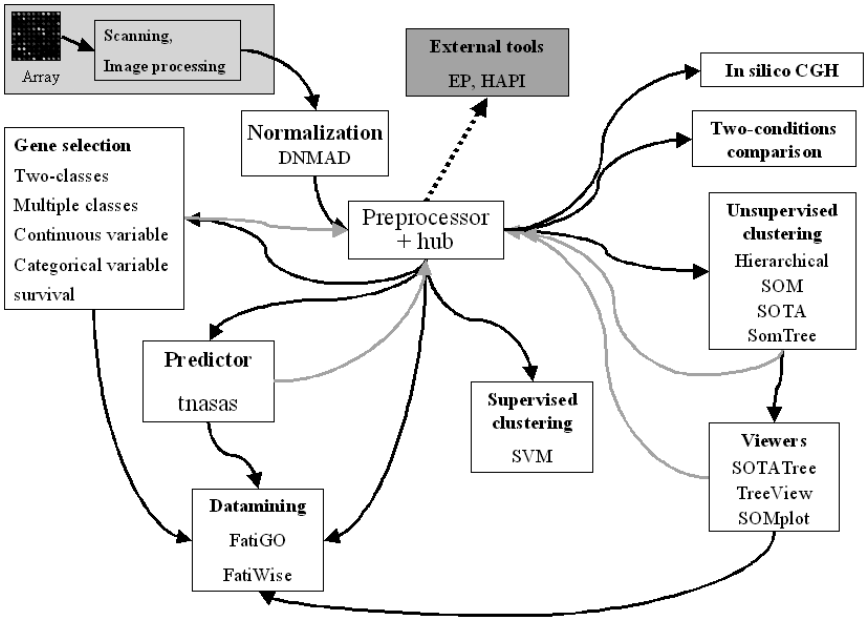

Figure 1. The pipeline of microarray data analysis. After the operations of image processing are performed (grey box on top left), the data enters the pipeline through the data normalization module and after normalization is sent to the pre-processor. Then, depending on the type of analysis the user needs to perform, these can be sent to different modules that implement different tools. External connectivity to other modules can be achieved from the preprocessor.

provided by web tools means that remote users can access resources without the hardware support for heavy calculations that are made on the server side, irrespective of the computer platform used.

GEPAS—A BRIEF DESCRIPTION

GEPAS incorporates several interconnected methods (Figure 1), implemented as individual modules, that allows to input data from the scanner, normalize them (currently normalization is only implemented for cDNA microarrays), performing preprocessing of data (log-transformations, standardizations, imputation of missing values, etc.) (21), and also provides different types of analysis including:

- Unsupervised clustering, comprising different hierarchical and non-hierarchical methods such as aggregative clustering (22), SOTA (23,16), SOM (24) (which implements a web interface to the SOM_PACK, http://www.cis.hut.fi/research/som_lvq_pak.shtml) and SOM-Tree (20), a mixture of SOM and aggregative clustering.
- Differential gene expression analysis, which involves finding genes showing significant differences in two or more experimental conditions or correlated to another phenotypic trait or experimental condition independent of the expression values (e.g. drug dosages, survival, level of a metabolite, etc.). This module, called Pomelo, is a tool that has been designed to address the problem of multiple testing when searching for differentially expressed genes. We have implemented four methods to account for multiple testing; two of them control the Family Wise Error Rate (25) and two others control de False Discovery Rate (25,27). These methods can be applied to five different statistical tests: the t-test (to compare expression between two conditions), ANOVA (Analysis of Variance, to compare expression between two or more conditions), linear regression (to examine if the expression of genes is related to variation in a continuous variable, e.g. expression levels of a given metabolite), survival analysis [to examine if gene expression is related to patients' survival (28)] and Fisher's exact test for contingency tables (when both the dependent and independent variables are categorical).
- A module for supervised classification based on the powerful methodology of Support Vector Machines [SVMs http://www.csie.ntu.edu.tw/~cjlin/libsvm (29,30)]. It allows the training of the model and further use for classification. SVMs have been applied successfully to classification problems in microarray data for both genes (30) and experiments (31).
- A module for data mining, FatiGO (13), that allows finding significant asymmetrical distributions of GO terms between groups of genes. This constitutes an extremely useful tool for exploring the biological meaning of the groups or arrangements of genes found by using the previous methods.
- A module for two-conditions comparison, which is, essentially, a viewer for comparing pairs of experiments.

In addition, new tools are included in this second version, which are described in the next section.

NEW TOOLS IN GEPAS

Since the original version of GEPAS, a number of new tools have been implemented to increase data analysis capabilities. To close the gap between the raw data, as obtained from the scanner, and the first step of data processing, a normalization module, DNMAD, has been implemented. There is a growing interest in obtaining predictors of clinical outcomes such as disease, resistance to therapies, life expectancy, etc. To obtain genes or combinations associated to these traits we have included Tnasas (for 'This is not a substitute for a statistician'), a module for building predictors. Translating the arrangements or clusters of data obtained upon the application of the different methods (clustering, gene selection, etc.) into actual biological meaning, requires data-mining tools (32). FatiGO (FAst Transference of Information using Gene Ontology) (13,14), a tool for extracting significant functional terms, obtained from Gene Ontology (9), has been fully integrated in GEPAS and, additionally, new databases and functionalities have been added to this tool. With the availability of accurate human and mouse genome maps, the possibility of gaining a genome-wide picture of the transcriptional process is within reach. In many cases, alterations in gene expression are due to amplifications or deletions that are evident when the altered genes are plotted on their genomic locations. We have developed InSilicoCGH, a tool that maps gene expression values over the chromosomes (L. Conde, J.C. Cigudosa and J. Dopazo, submitted).

Knowledge filtering

This module takes a matrix of expression values, extracts the gene identifiers and sends a query to the corresponding database. At this stage, queries are made through the FatiGO engine. It returns an output that allows the selection of genes based on their gene ontology annotations. Based on the labels corresponding to the GO terms, genes can be highlighted in the cluster visualization tools. In this way, functional and biological information can be mapped on the clustering results.

Normalization

DNMAD is a web interface to aid the normalization of cDNA microarray data and constitutes the real starting point for proper analysis of microarray data. The method implemented is the print-tip loess as explained in (33,34) . Essentially, the objective of the normalization is to adjust for effects explained by variations in the technology rather than the actual biology. Specifically, the algorithm tries to adjust for differences in the red and green labeling caused, for example, by differences in the binding of the labels. Since these differences can be related to which print-tip printed each spot, the adjustment is carried out, generally, for each print-tip separately. Thus, the basic normalization is based on a *print-tip loess*, which fits a robust local regression to the relation between M (difference in log intensities) and A (the 'average' staining). The normalized M-value is the original one minus the loess fitted one, and thus should correct for spatial effects (as reflected by print-tips) and for effects related to intensity. Different diagnostic plots (before and after normalization) are available such as *MA-plots*, which show the relationship between A (the 'average signal') and M [the $\log_2$ differential ratio: $\log_2(R/G)$]; *box plots*, which consists of the median, the upper and lower quartiles, the range, and individual extreme values (http://www.bioconductor.org), and a set of diagnosis images to help in the detection of problems due to scanner adjustment,

positional effects, etc., which include *histograms of the raw pixel intensities*, which provide the logarithm of the red and green mean foregrounds and images of the arrays with red and green background as well as unnormalized and normalized *M*. Finally, it is essential to check that the normalization is working in terms of *scale* (approximate variance).

Since many arrays can be normalized at the same time, DNMAD provides *box plots intra-array* to help assess if there are differences in scale among print-tips within array, and *box plots inter-array* to assess if there are differences in scale among arrays. If there are large differences in scale among arrays, a further normalization for scale can be performed. This is scarcely required, and introduces additional noise.

Additional options include the possibility of using spot's flags, optional use of background subtraction and the possibility of using global loess (instead print-tip).

Tnasas, a predictor-building tool

Tnasas is a tool that implements a simple, yet often effective way of building class predictors from microarray data. We use three basic types of predictors: nearest neighbour, diagonal linear discriminant analysis and support vector machines. The user can select the type of predictor. Tnasas finds the number of genes (from a small set of possible numbers of genes) for building the predictor that yields the smallest cross-validated prediction error rate. In other words, Tnasas performs a simple form of 'variable selection' or 'feature selection'. The error rate, as it is computed taking into account the effect of gene selection, is not biased down by the 'selection bias' problem, as is so common in many microarray studies [e.g. (17,35)]. Moreover, Tnasas provides an honest (unbiased) estimate of the prediction error rate for the predictor built using this scheme of selecting the number of genes.

Currently our tool does not pretend to build 'the optimal predictor', but it can lead to reasonably effective predictors. These predictors have repeatedly been shown as good, or even better than, much more complicated algorithms and/or learning rules (36,37). Moreover, it has an important pedagogical value by highlighting the very serious underestimates of error rate that are caused by not taking the 'selection bias' problem into account. In addition, the set of predictors plus the simple mechanism for gene selection provide a straightforward and easy to use benchmark against some (overly) optimistic claims that occasionally are attached to new methods and algorithms. This is a particularly important feature, since many new predictor methods are being proposed in the literature often without careful comparisons with alternative methods; Tnasas can be used as a simple, effective way of comparing the performance of the newly proposed methods and can, itself, become a benchmarking tool.

FatiWise, the expanded FatiGO

The arrangements of genes based on their different behaviours under distinct experimental conditions (e.g. differential gene expression between diseased samples and controls, etc.) are a consequence of the biological roles that the genes are playing within the cell. To understand the biological processes operating throughout a given experiment, we developed a simple but powerful procedure to extract Gene Ontology terms that are significantly over- or under-represented in sets of genes. The procedure has been implemented as a web application, FatiGO (13), allowing for easy and interactive querying. FatiGO takes the multiple-testing nature of statistical contrast into account. FatiWise consists of an extension of the statistical framework, implemented in FatiGO, to other types of relevant biological knowledge rather than GO terms. FatiWise includes correspondence tables between genes and InterPro motifs (38), KEGG (10) pathways and SwissProt keywords. It can be used to study the distribution of genes belonging to different pathways within groups of genes. InterPro motifs allow the study of molecular functions of the genes based on criteria different from GO. In addition to function, InterPro entries account for structural and physical properties of proteins, increasing the scope of properties that can be studied. Something similar occurred with the keywords associated to SwissProt entries. In addition, the curation process of this database is known to be among the most rigorous ones.

InSilicoCGH

Alterations in the genome that lead to changes in DNA sequence copy number are a characteristic of solid tumours and are found in association with developmental abnormalities and/or mental retardation. CGH methodologies can be used to detect and map these changes. Recent improvements in the resolution and sensitivity of CGH have been possible through implementation of microarray-based CGH (array CGH) (4). The InSilicoCGH tool allows mapping the results of microarray hybridizations onto the chromosome coordinates . A number of different array platforms have been used for CGH measurements in mammalian genomes. The various approaches have employed large insert genomic clones, such as bacterial artificial chromosomes (BACs) (39), cDNA clones (40) and oligonucleotides for array spots. In any case, the tool retrieves the chromosomal coordinates of the probes in the array (irrespective of their nature—clones or BACs) and plots the hybridization values over the corresponding positions in the chromosome. Different identifiers for the probes are accepted by the program including Ensembl IDs, accession, EMBL accession, unigene codes, hugo names, refseq, BAC names, Ensembl's external IDs and internal CNIO IDs. The output provides three different views (Figure 2): *CGH*, which mimics a CGH representation by plotting in different colours over- and under-represented matches (Figure 2A); *lines*, which correspond to a bar graph of the hybridization values, plotted in the coordinate chromosomes (Figure 2B); and *karyotype*, that generates a representation with the appearance of a karyotype (Figure 2C) in which probes with hybridization values over a given threshold are mapped. There are different options for the representations, and plotting multiple arrays is possible. Figure 2D shows the magnifier tool, which provides a magnified view of any part of the representation.

The tool can be used for visualizing the hybridization of mRNA or genomic DNA on the chromosomal positions. The connectivity provided by GEPAS allows mapping not only complete arrays on the genome, but also clusters of coexpressing genes from the tree viewer.

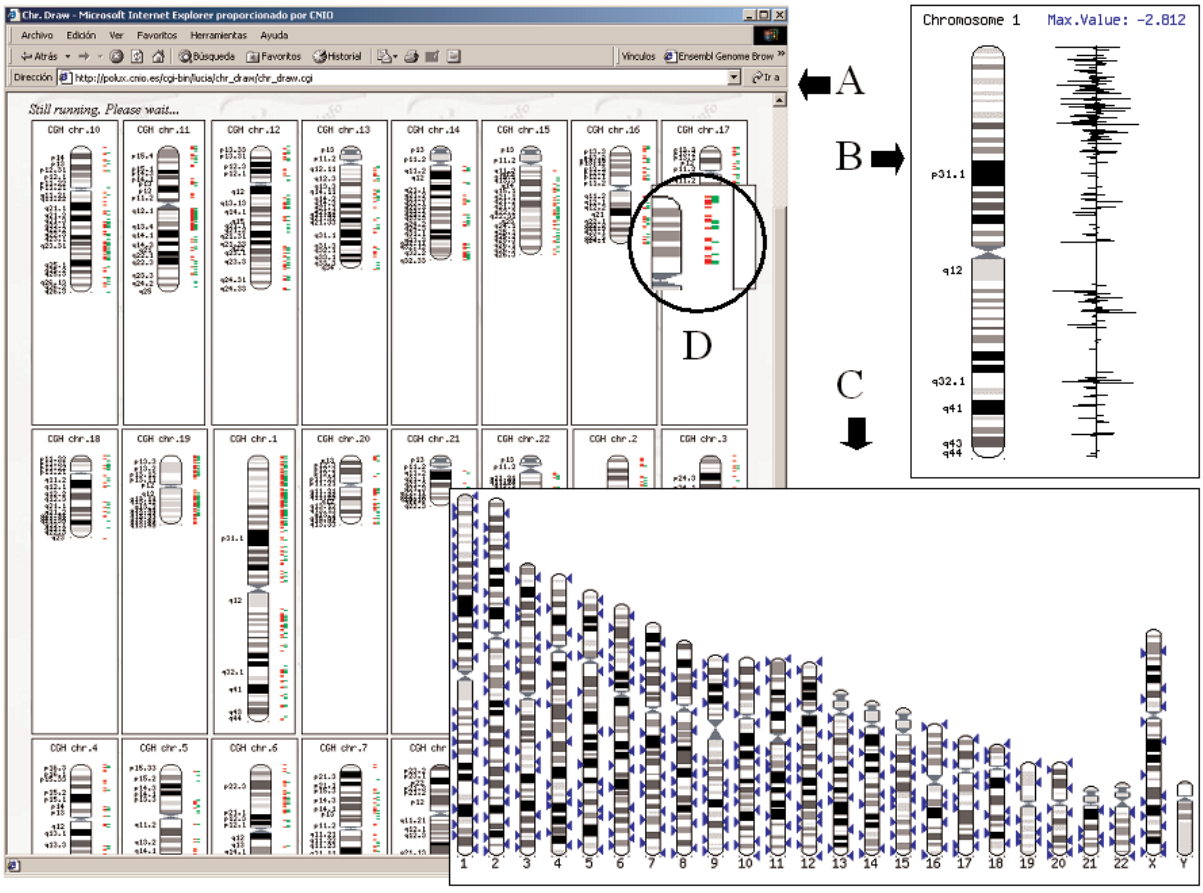

Figure 2. Results of the InSilicoCGH module. Three different views are available: (**A**) *CGH*, which mimics a CGH representation by plotting in different colours overrepresented and underrepresented matches; (**B**) *lines*, which correspond to a bar graph of the hybridization values, plotted in the coordinate chromosomes; and (**C**) *karyotype*, that generates a representation with the appearance of a karyotype in which probes with hybridization values over a given threshold are mapped. There are different options for the representations, and plotting multiple arrays is possible. (**D**) Within the circle the effect of the magnifier tool can be seen. The magnifier allows to obtain a magnified view a any part of the representation.

COMPARISON TO OTHER WEB-BASED TOOLS FOR MICROARRAY DATA ANALYSIS

Despite the growing number of tools available for the analysis of gene expression data, only a few are web based. Obviously, in web-based solutions graphical interface and interactivity are, to some extent, sacrificed. Nevertheless, this is in exchange for a number of advantages that include cross-platform usage and availability of the calculation power of the server irrespective of the client used.

The website of Y. F. Leung is probably among the most complete ones and, more importantly, it is updated (http://ihome.cuhk.edu.hk/~b400559/arraysoft.html). The site has a comprehensive list of programs for microarray data analysis. Another important source of information is the Stanford Microarray Database page (http://genome-www5.stanford.edu/). Among the tools listed are several web-based applications:

Cleaver (http://classify.stanford.edu/) a web tool which allows classification (discriminant analysis) and clustering by *K*-means.

ENGENE (http://www.engene.cnb.uam.es/). This is an exploratory tool that includes several clustering methods as well as visualization procedures.

EP (http://www.ebi.ac.uk/expressionprofiler/) (1). A set of tools for clustering, analysis and visualization of gene expression and other genomic data.

GEDA (http://bioinformatics.upmc.edu/GE2/GEDA.html). A gene-expression data-analysis web application with a variety of preprocessing steps, tests for differentially expressed genes and clustering algorithms.

CTWC (http://ctwc.weizmann.ac.il/). Performs coupled Two-Way Clustering (42), a biclustering method for finding subsets of genes and samples of interest.

INCLUSive (http://www.esat.kuleuven.ac.be/%7Edna/Biol/Software.html). A suite of web-based tools that is aimed at the automatic multistep analysis of microarray data (clustering and motif finding). Currently, adaptive quality-based clustering, retrieval of upstream sequences and the motif sampler are accessible from this website.

R Cluster (http://genomics.biochem.uci.edu/cgi-bin/genex/rcluster/index.cgi). Web interface to a collection of clustering routines written in the R statistical programming language.

In addition, there are other web-based tools for specialized tasks such as normalization and functional annotation of

microarray experiments:

OntoTools (http://vortex.cs.wayne.edu/Projects.html). OntoTools is composed of Onto-Express, translate differentially regulated genes into functional profiles; Onto-Compare, comparisons of any sets of commercial or custom arrays; Onto-Design, select genes that represent given functional categories; and Onto-Translate, translate lists of accession numbers, UniGene clusters and Affymetrix probes into one another.

Multi Microarray Normalization (http://genome1.beatson.gla.ac.uk/Rweb/anova.html). An ANOVA based normalization of dye-swapped experiment, taking pin-tip effect into account.

SNOMAD (http://pevsnerlab.kennedykrieger.org/snomadinput.html). Standardization and NOrmalization of Micro Array Data is a collection of algorithms directed at the normalization and standardization of DNA microarray data.

Many of the tools focus on clustering or, in general, on unsupervised classification methods and only one of them (GEDA) provides support for differential gene expression analysis. Features such as a search method for putative transcription factor binding sites are available in EP (only for yeast) and INCLUSive. Normalization and functional annotation can only be found as separate applications. The trend is similar for stand-alone tools, although there are more possibilities for differential gene expression analysis and other types of analysis. Nevertheless, none of the programs or packages provides tools for obtaining predictors. The number of tools, the degree of integration and the scope makes GEPAS one of the most complete packages for microarray data analysis, even when compared with stand-alone packages or with commercial solutions.

In addition, GEPAS modules can be invoked from other web resources and vice versa. This allows other designers of web tools to use partial or full GEPAS resources. At present, GEPAS can send data files, in the proper format, to Expression Profiler (41), and to HAPI, a data-mining tool based on hierarchies of MESH terms (43).

Usage of GEPAS

In addition to remote web usage, GEPAS is freely distributed upon request. If remote usage is a problem because of excessive Internet traffic, GEPAS can be locally installed. The requirements are simple: an apache web server, linux OS, PERL, R and some free packages (see GEPAS information http://gepas.bioinfo.cnio.es/mirrors.html). Also, the source code of the modules developed by us are publicly available in our downloads web page (see http://bioinfo.cnio.es/downloads/).

Our records show that, since April 2003, GEPAS has been used more than 24 000 times, with a daily average of more than 80 uses. The approximate distribution of users is as follows: 25% Spain, 15% US (domains .edu, .com, and .net), 10% France, 5% UK, and lower percentages of 1–2%, depending on the month, are due to users from Japan, The Netherlands, Italy, Germany, Portugal and Chile mainly. FatiGO usage, not included in the above figures, accumulates itself more than 11 000 uses, with a daily average of more than 50 uses.

CONCLUSIONS

Despite the growing number of programs and packages available for microarray data analysis, there are still many aspects of data analysis with poor or incomplete coverage. Most of the software available for microarray data analysis focusses on unsupervised cluster methods that, in many cases, are used for inadequate purposes (17). Since the first release (16), GEPAS has evolved to cope with new challenges arising in the field of microarray data analysis.

Connectivity is also a problem: different tools perform different tasks that constitute consecutive steps of analysis. This causes problems with input/output formats. GEPAS provides the user with an integrated environment in which modules for different types of analysis, which respond to real analysis demands, can be found.

From a technical point of view, GEPAS has been designed with the intention of taking full advantage of the web properties: connectivity, cross-platform and remote usage. The modular architecture allows the addition of new tools and facilitates the connectivity of GEPAS from and to other web-based tools.

ACKNOWLEDGEMENTS

Our special thanks to Amanda Wren for revising the English of the manuscript. F.A. is supported by grant BIO2001-0068 from MCyT, A.M. is supported by an IBM fellowship, L.C. is supported by a fellowship from the FIS (grant PI020919). R.D.U. is supported by a Ramón y Cajal research contract from the MCyT. This work is partly supported by grants from Fundación Ramón Areces and Fundació La Caixa.

REFERENCES

1. Eisen,M., Spellman,P.L., Brown,P.O. and Botsein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.
2. van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T., Schreiber,G.J., *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
3. Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.*, **8**, 816–824.
4. Albertson,D.G. and Pinkel,D. (2003) Genomic microarrays in human genetic disease and cancer *Hum. Mol. Genet.*, **12**, 145R–152R.
5. Oliveros,J.C., Blaschke,C., Herrero,J., Dopazo,J. and Valencia,A. (2000) Expression profiles and biological function. *Genome Inform.*, **10**, 106–117.
6. Raychaudhuri,S., Schutze,H. and Altman,R.B. (2002) Using text analysis to identify functionally coherent gene groups. *Genome Res.*, **12**, 1582–1590.
7. Pavlidis,P., Lewis,D.P. and Noble,W.S. (2002) Exploring gene expression data with class scores. *Pac. Symp. Biocomput.*, **7**, 474–485.
8. Blaschke,C., Hirschman,L. and Valencia,A. (2002) Information extraction in molecular biology. *Brief. Bioinform.*, **3**, 154–165.
9. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
10. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

11. Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*., **4**, R7.
12. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002). Profiling gene expression using onto-express. *Genomics*, **79**, 1–5.
13. Al-Shahrour,F., Díaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms to groups of genes. *Bioinformatics*, **20**, 578–580.
14. Al-Shahrour,F., Herrero,J., Mateos,A., Santoyo,J., Diaz-Uriarte,R. and Dopazo,J. (2003) Using Gene Ontology on genome-scale studies to find significant associations of biologically relevant terms to group of genes. In Neural Networks for Signal Processing XIII. IEEE Press. New York, pp. 43–52.
15. Slonim,D.K. (2002) From patterns to pathways: gene expression data analysis comes of age. *Nature Genet*., **32** (Suppl), 502–508.
16. Herrero,J., Al-Shahrour,F., Díaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res*., **31**, 3461–3467.
17. Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst*., **95**, 14–18.
18. Schena,M., Shalon,D., Heller,R., Chai,A., Brown,P.O. and Davis,R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci., USA*., **93**, 10614–10619
19. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M, Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol*., **14**, 1675–1680
20. Herrero,J. and Dopazo,J. (2002) Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. *J. Proteome Res*., **1**, 467–470.
21. Herrero,J., Díaz-Uriarte,R. and Dopazo,J. (2003) Gene expression data preprocessing. *Bioinformatics*, **19**, 655–656.
22. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. W.H. Freeman, San Francisco, CA.
23. Dopazo,J. and Carazo,J.M. (1997) Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. *J. Mol. Evol*., **44**, 226–233.
24. Kohonen,T. (1997) Self-organizing maps. Springer-Verlag. Berlin.
25. Westfall,P.H. and Young,S.S. (1993) *Resampling-based Multiple Testing: Examples and Methods for P-value Adjustment*. John Wiley & Sons, New York.
26. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
27. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat*., **29**, 1165–1188.
28. Klein,J.P. and Moeschberger,M. L. (1997) *Survival Analysis*. Springer-Verlag. New York.
29. Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.
30. Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl Acad. Sci., USA*, **97**, 262–267.
31. Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
32. Díaz-Uriarte,R., Al-Shahrour,F. and Dopazo,J. (2003) Use of GO terms to understand the biological significance of Microarray Differential Gene Expression Data. In Johnson,K.F. and Lin,S.M. (eds), *Microarray Data Analysis III. Kluwer Academic*. pp. 233–247.
33. Yang,Y.H., Dudoit,S., Luu,P., Lin,D.M., Peng,V., Ngai,J. and Speed,T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., **30**, e15.
34. Smyth,G.K., Yang,Y.H. and Speed,T.P. (2003) Statistical issues in microarray data analysis. In Brownstein,M.J. and Khodursky,A.B. (eds), *Functional Genomics: Methods and Protocols, Methods in Molecular Biology*. Humana Press, Totowa,NJ, Vol. 224, pp. 111–136.
35. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci., USA*, **99**, 6562–6566.
36. Dudoit,S., Fridlyand,J. and Speed,T.P. (2002) Comparison of discrimination methods for the classification of tumors suing gene expression data. *J. Am. Stat. Assoc*., **97**, 77–87.
37. Romualdi,C., Campanaro,S., Campagna,D., Celegato,B., Cannata,N., Toppo,S. Valle,G. and Lan franchi,G. (2003) Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum. Mol. Genet*., **12**, 823–836.
38. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
39. Snijders,A.M., Nowak,N., Segraves,R., Blackwood,S., Brown,N., Conroy,J., Hamilton,G., Hindle,A.K., Huey,B., Kimura,K. *et al*. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet*., **29**, 263–264.
40. Pollack,J.R., Perou,C.M., Alizadeh,A.A., Eisen,M.B., Pergamenschikov,A., Williams,C.F., Jeffrey,S.S., Botstein,D. and Brown,P.O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet*., **23**, 41–46.
41. Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett*., **480**, 17–24.
42. Getz,G. and Domany,E. (2003) Coupled two-way clustering server. *Bioinformatics*, **19**, 1153–1154.
43. Masys,D.R., Welsh,J.B., Fink,J.L., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **7**, 319–326